ARTICLE

DISPARITIES IN SENSE OF COMMUNITY: TRUE RACE DIFFERENCES OR DIFFERENTIAL ITEM FUNCTIONING?

Donna L. Coffman
The Methodology Center, The Pennsylvania State University

Rhonda BeLue
The Methodology Center and Department of Health Policy Administration, The Pennsylvania State University

The sense of community index (SCI) has been widely used to measure psychological sense of community (SOC). Furthermore, SOC has been found to differ among racial groups. Because different ethnic groups have different cultural and historical experiences that may lead to different interpretations of measurement items, it is important to know whether the instrument used to measure the construct of interest has equivalency in measurement across groups or if the instrument exhibits differential item functioning (DIF). Examining DIF in the SCI helps assure that subgroup comparisons identify true differences in SOC between Blacks and Whites. Although we did not find DIF between races, we did find that the SCI question, "I feel at home in my neighborhood," was a more reliable measure of SOC for Whites than for Blacks. In other words, this item has less measurement error for Whites than for Blacks. Therefore, differences on the SCI may be attributable to true differences in SOC between races rather than DIF. © 2009 Wiley Periodicals, Inc.

Sense of community (SOC) is a psychological construct that has been widely applied as an essential principle in the fields of community based research and community psychology (Chipuer & Pretty, 1999; Chavis & Pretty, 1999). Initially proposed by

Preparation of this paper was supported by National Institute on Drug Abuse Center grant P50 DA10075 and National Institute on Drug Abuse Training grant T32 DA017629-01A1. We would also like to thank Drs. Margaret Hargreaves, Celia Larson, and David Schlundt from Nashville REACH 2010 for facilitating the parent project.
Correspondence to: Donna L. Coffman, The Pennsylvania State University, 204 E. Calder Way, Ste. 400, State College, PA 16801. E-mail: dlc30@psu.edu

JOURNAL OF COMMUNITY PSYCHOLOGY, Vol. 37, No. 5, 547–558 (2009)
Published online in Wiley InterScience (www.interscience.wiley.com).
 DOI: 10.1002/jcop.20312

Sarason (1974, 1986) and refined by later researchers (McMillan & Chavis, 1986), McMillan (1996) defines SOC as "a *spirit* of belonging together, a feeling that there is an authority structure that can be *trusted*, an awareness that *trade*, and mutual benefit come from being together, and a spirit that comes from shared experiences that are preserved as *art*" (p. 315).

The sense of community index (SCI; McMillan & Chavis, 1986) has been widely used to measure SOC. The SCI has been used to ascertain psychological perceptions related to community belonging in both geographic and relational representations (e.g., among employees within an institution) of communities (Bronsky, O'Campo, & Aronson, 1999; Klein & D'Aunno, 2006).

SOC is operationalized using the following four constructs: (a) membership—the feeling that one is a member of the community; (b) influence—the belief that one can have influence over or be influenced by community issues; (c) reinforcement of needs—the belief that members of the community desire the same things and have the same values and needs; and (d) shared emotional connection—the feeling that those residing in the community get along and desire to live in the community. Previous research has examined the factor structure of the SCI and the conclusions have been mixed. Chipuer and Pretty (1999) tested the factor structure of the SCI and concluded that it can be used as one-factor instrument. Their study included four different samples, and although the factor analyses showed some support for four factors, the results were not consistent across samples. More recently, Peterson, Speer, and McMillan (2008) reported on a factor analysis of the Brief Sense of Community Scale (BSCS) and found support for four factors. This scale consists of eight items based on the four-factor theory of SOC and the response option format is a 5-point Likert scale.

Results from reliability analyses have similarly been mixed. It should be noted that if the SCI is multidimensional, then it is not appropriate to compute coefficient alpha as a measure of the reliability of the entire scale. Of course, coefficient alpha is also specific to a particular sample. Although the primary purpose of the present study is to examine the measurement properties of the SCI across racial groups, we assess dimensionality and reliability of the SCI as a preliminary step.

The SCI has been used across a wide variety of race and ethnic groups and across multiple contexts. Perceived sense of community and belonging has been found to differ among racial/ethnic groups (BeLue, Taylor-Richardson, Lin, McClellan, & Hargreaves, 2006; Faircloth & Hamm, 2005; Clark, Colantonio, Rhodes, & Escobar, 2008). Because different ethnic groups have different cultural and historical experiences that may lead to different interpretations of measurement items, it is important to know whether the instrument used to measure the construct of interest has equivalency in measurement across groups (Rameriz, Ford, Stewart, & Teresi, 2005). Differential item functioning (DIF) can contribute to observed differences in outcome indicators that are attributable to item non-invariance across groups as opposed to true between-group differences. Specifically, DIF refers to the situations in which the probability of endorsement of a particular item differs by subgroup characteristics after adjusting for trait level, or latent variables of interest (Cole, 1999). Identifying and correcting DIF in SOC can help assure that subgroup comparisons identify true differences in SOC among race groups, especially when examining the relationship between correlates of SOC that are known to differ by race/ethnicity. To our knowledge, previous research has not examined DIF among racial groups.

Purpose of the Present Study

The purpose of the present study is to investigate the psychometric properties of the SCI in a diverse community sample. Specifically, we investigate whether the SCI is invariant between Blacks and Whites to determine whether the index measures the same construct in both groups. This analysis of measurement invariance (Meredith, 1993) is essential to understanding how SOC may truly vary among racial groups (Thissen, Steinberg, & Gerrard, 1986).

Finally, we describe the practical implications (often referred to as differential impact) and offer suggestions to correct for DIF so that the groups may be meaningfully compared on the outcome of interest.

METHOD

Data Collection Methods and Sample

The sample. A random household survey designed to assess socio-demographic information, health status, health behavior and correlates of chronic diseases among a diverse sample was collected in a midsized Southern city in 2005. A total of 1,463 interviews were completed from a sample of 9,576 residential telephone numbers. Phone numbers were purchased from an independent survey sampling company.

The survey. The data for this study was collected using computer assisted telephone interviewing (CATI) technology. The CATI system randomly selects sample telephone numbers to be dialed and allows the interviewers to conduct the interview and enter respondent answers, simultaneously. The CATI system controls queuing of questions and scheduling of callbacks. Telephone interviews were conducted in the first 6 months of 2005. Interviews were conducted by trained interviewers on weekday evenings and Saturdays. Interviewers made up to five attempts to reach each potential respondent. The sample for the present analyses consists of 1,428 (588 Whites and 840 Blacks) individuals. For additional information on data collection methods for this study, please see the previously published paper by BeLue et al. (2006).

Measures

Socio-demographics. Socio-demographic characteristics measured in this study include respondent age, gender, marital status, and self-reported ethnicity/race. Level of education was classified as one of the following: less than high school, high school or greater.

Sense of community. Included in the survey was the SCI (Chavis, 1986; 1990) that was designed to measure SOC in adults in a neighborhood. Higher scores indicate a stronger SOC. The response scale was dichotomous (agree or disagree) and there were

Table 1. Sense of Community Index and Factor Loadings

Item	*Factor loading*	*Biserial correlations*	*Facility*
1) I think my neighborhood is a good place to live.	.84	0.656	0.918
2) People in this neighborhood believe the same things are important.	.84	0.694	0.842
3) My neighbors and I want the same thing.	.83	0.694	0.846
4) I can recognize many people who live in my neighborhood.	.55	0.494	0.841
5) I feel at home in my neighborhood.	.86	0.614	0.947
6) Very few of my neighbors know me.	.53	0.500	0.697
7) I care about what my neighbors think of my actions.	.54	0.541	0.764
8) I have almost no influence over what this neighborhood is like.	.49	0.505	0.638
9) If there is a problem in this neighborhood, the people here can solve it.	.75	0.664	0.826
10) It is very important to me to live in this neighborhood.	.78	0.703	0.726
11) I expect to live in this neighborhood for a long time.	.67	0.590	0.790
12) People around here are willing to help their neighbors.	.79	0.675	0.892

12 items so the total SOC score ranged from 0 to 12 (1 = *agree*, 0 = *disagree*). The items are given in Table 1. Questions 6 and 8 were reverse scored.

Statistical Analysis

Chi-square tests were performed to assess racial differences in demographic variables. To assess dimensionality, we took an item response theory (IRT) approach using full information item factor analysis (Bock, Gibbons, & Muraki, 1988) with marginal maximum likelihood estimation (Bock & Aitkin, 1981). Following dimensionality assessment, we fit several IRT models and chose one based on the Akaike information criterion (AIC; Akaike, 1987), Bayesian information criterion (BIC; Schwarz, 1978), and log-likelihood nested tests. Finally, we assessed whether the items function differentially between Blacks and Whites.

IRT is a set of methods and models to assess questionnaire items. When item responses are dichotomous, as they are in the present analysis, there are several models from which to choose. These include the one-parameter logistic (1PL) model, also known as the Rasch model, the two-parameter logistic (2PL) model, and the three-parameter logistic (3PL) model. The 1PL model includes a single parameter whereas the 2PL model includes two parameters. The 3PL model includes a third parameter that is interpreted as a guessing parameter when analyzing educational tests. Therefore, the 3PL model is not appropriate for the present analysis. The single parameter (denoted b) of the 1PL model is a location parameter and is known as the difficulty parameter in the educational testing literature. This parameter represents the value on the latent distribution at which the probability of endorsing the item is .5. The slopes, or regression weights, of the items on the latent construct, SOC, are assumed to be equal across all items. The second parameter (denoted a) in the 2PL model is the slope, or regression weight, of an item on the latent construct, and it is usually called the discrimination parameter in the educational literature. This parameter represents the degree to which an item discriminates among those with a similar SOC and quantifies the relation between the item response and the latent construct. Thus, unlike the 1PL model, the 2PL model does not assume that all items are related to the latent construct to the same degree (Thissen & Steinberg, 1988). The slope parameters are directly related to factor loadings for a unidimensional 2PL model (Takane & DeLeeuw, 1987).

IRT proposes that reliability is not the same across the entire range of the latent construct. Information is the IRT extension of the concept of reliability. More information implies less measurement error and, therefore, greater measurement precision. The results of an IRT analysis are best described graphically. In particular, item characteristic curves, item information curves, and test information functions provide for examination of individual items and the overall scale. The test information function is the sum of the item information curves. The item characteristic curves describe the relations of the item responses to the underlying latent construct, SOC.

IRT is a promising tool for evaluating and establishing measurement equivalence between racial and ethnic groups. DIF is assessed by examining whether the item parameters differ across groups (Thissen, Steinberg, & Wainer, 1988; 1993). The statistical test for DIF is distributed as a chi-square statistic and tests the null hypothesis that all item parameters are equal across groups. If the null hypothesis that both the *a* and the *b* parameters are equal across groups is tested, then the test statistic is compared with a chi-square distribution with 2 degrees of freedom. If the null hypothesis that the location parameter is equal across groups is tested, then the test statistic is compared with a chi-square distribution with 1 degree of freedom. If this hypothesis is rejected, then the item is said to exhibit *b*—DIF. Similarly, if the null hypothesis that the slope parameter is equal across groups is tested, then the test statistic is compared with a chi-square distribution with 1 degree of freedom. If this hypothesis is rejected, then the item is said to exhibit *a*—DIF.

RESULTS

Demographics

The respondents for this study had a mean age of 55 ($SD = 17.6$), were most likely to be female (69%), and were majority Black (59%) and single (59%). Approximately 85% of the survey respondents have completed high school or more Chi-square tests indicated that Blacks were less likely to be married and less likely to have completed high school or greater. These demographics, as well as others, are given in Table 2.

Assessing Dimensionality

First, we examined whether the SCI measures a single dimension or whether, as suggested by some previous research, it measures four dimensions. The item factor analysis was carried out in TESTFACT. Based on a scree plot (shown in Figure 1) of the eigenvalues and interpretability, we retained one factor. Thus, the items measure a single construct. The standardized factor loadings are given in Table 1. Also presented in Table 1 are the item facilities and biserial correlations. The item facility is the proportion of respondents who endorsed an item. Thus, 91.8% of respondents agreed with the first item that their neighborhood is a good place to live. The biserial correlation is a measure of association between performance on an item and performance on the overall scale. Thus, for example, item number 4 is less associated with the overall scale than item 1.

Table 2. Sample Demographics

	All N = 1428	Blacks N = 840 59% of total	Whites N = 588 41% of total
		% in each category	
Marital status[a]			
Married	40.8	30.6	55.3
Single	59.2	69.4	44.7
Gender			
Male	30.7	28.2	34.2
Female	69.3	71.1	65.8
Education[a]			
Less than high school	15.4	17.4	12.5
High school or greater	84.6	83.6	88.5
Age	$M = 54.5$ $SD = 17.61$	$M = 55.07$ $SD = 17.72$	$M = 53.71$ $SD = 17.41$

[a]Indicates $p < 0.05$ as determined by a chi-square test.

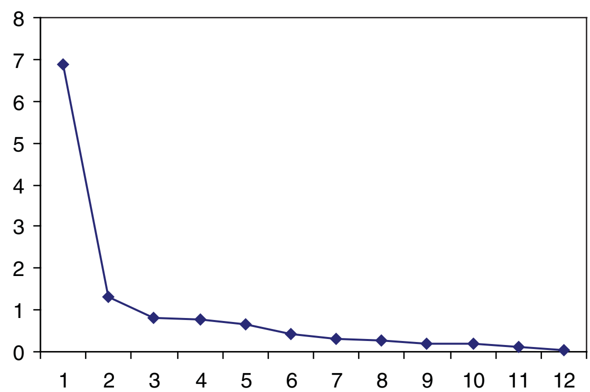

Figure 1. Scree plot of eigenvalues from factor analysis.

Choosing an IRT Model

Using the latent trait modeling (LTM) package in R (Rizopoulos, 2006), we fit a 1PL model with the slope parameter fixed to 1, a 1PL model with the slope estimated but equal across all items, and a 2PL model. The 2PL model was chosen on the basis of the AIC, BIC, and the log-likelihood nested tests. The difference between the log-likelihood for two nested models is chi-square distributed with degrees of freedom equal to the difference between the degrees of freedom for the two models. These fit measures for each of the models are given in Table 3. The parameter estimates for the 2PL model are given in Table 4 under the heading "Entire Sample."

Figures 2, 3, and 4 present the item characteristic curves, item information curves, and the test information function, respectively, for the entire sample. The *x*-axis indicates the position on the latent construct, SOC, on a standardized scale with mean zero and standard deviation of 1. The *y*-axis for the item characteristic curve indicates the probability of endorsing the item. For this sample, the item characteristic curves in Figure 2 show that those with a slightly below average SOC have a 50% likelihood of

Table 3. Measures of Fit for Models Fit to Entire Sample

Model	*AIC*	*BIC*	*Log likelihood*
1PL w/constraint	12668.76	12731.92	−6322.378
1PL no constraint	12371.47	12439.90	−6172.735
2PL	12157.40	12283.74	−6054.702

Note. AIC indicates Akaike information criterion; BIC = Bayesian information criterion.

Table 4. Parameter Estimates for the 2PL Model for the Entire Sample and by Race

	Entire sample		*White*		*Black*	
Item	*Slope*	*Location*	*Slope*	*Location*	*Slope*	*Location*
1	3.11	−1.56	2.84	−1.89	3.18	−1.42
2	3.01	−1.10	2.81	−1.24	3.06	−1.03
3	2.99	−1.12	2.55	−1.30	3.23	−1.04
4	1.26	−1.67	1.39	−1.53	1.21	−1.76
5	3.61	−1.74	4.78	−1.77	3.31	−1.74
6	1.14	−0.91	1.27	−0.86	1.08	−0.94
7	1.21	−1.23	0.98	−1.63	1.33	−1.06
8	1.03	−0.67	1.30	−0.69	0.87	−0.67
9	2.23	−1.13	1.94	−1.19	2.49	−1.11
10	2.43	−0.74	2.54	−0.84	2.31	−0.70
11	1.74	−1.11	1.64	−1.29	1.79	−1.03
12	2.60	−1.44	2.20	−1.85	2.86	−1.25

endorsing or agreeing with the items. The exact values are given in Table 4 under location. As is also evident from Figure 2, the slopes of the item characteristic curves are steep. The actual values are given in Table 4 under slope. Those items with shallow slopes are the same items with lower factor loadings in Table 1. Most of the items are strongly related to the latent construct.

As is evident in the item information curves of Figure 3, the items provide more information and, therefore, more precision in the below average range for SOC. Thus, the addition of items to the SCI that provides information in the above average range for SOC would enable precise measurement of the construct across the entire range of scores. This type of detailed item analysis is not possible in classical test theory from which we learn only that the coefficient alpha for the entire sample was .82 indicating an acceptable reliability for the total scale.

DIF Analysis

First, we fit the 2PL model separately in each group. Then we tested constraints for the item parameters across the two groups. The parameter estimates for the 2PL model fitted separately for Blacks and Whites are presented in Table 4 under the respective headings. Examination of Table 4 reveals that some items are more strongly related to the latent construct for Blacks, while other items are more strongly related to the latent construct for Whites. The item characteristic curves, item information curves, and test information function are given in Figures 2, 3, and 4, respectively, for Blacks and Whites.

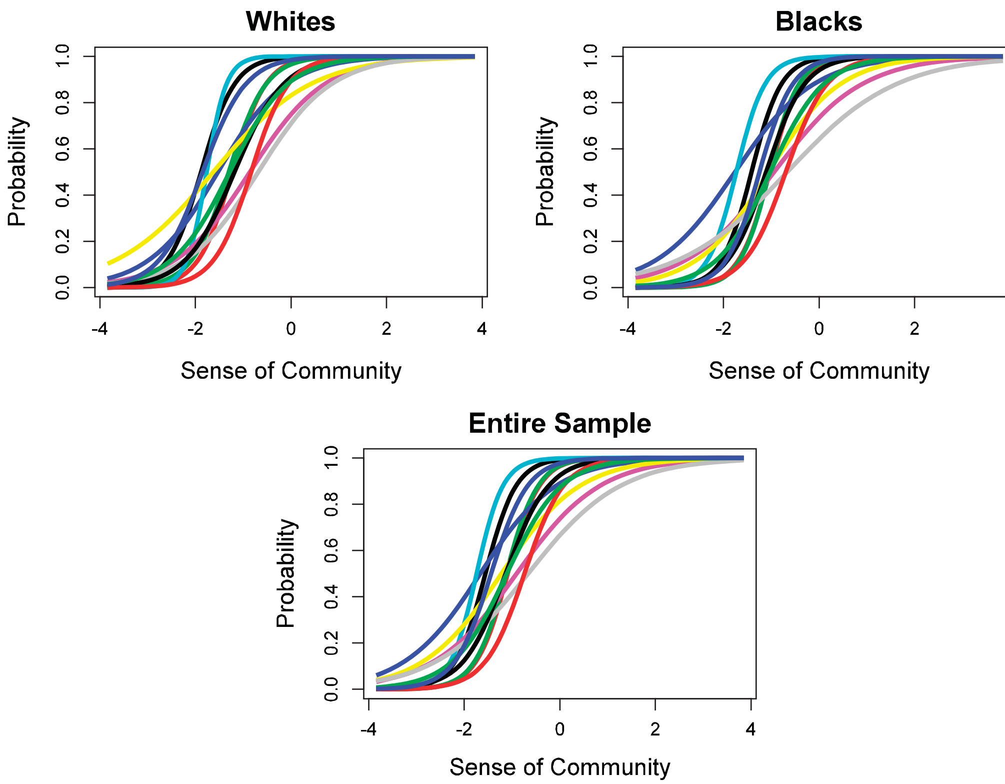

Figure 2. Item characteristic curves by race and for entire sample.

For Blacks, many of the items provide a similar amount of information, but for Whites, one item in particular (Item 5) provides more information than the others. However, the test information functions for both groups are similar in that the items provide more information and, therefore, more precision in the below average range for SOC. We now test these differences statistically.

At the present time, neither the LTM R package, which is free, or TESTFACT perform statistical tests of DIF. Therefore, we used the free program IRTLRDIF (which may be obtained at http://www.unc.edu/~dthissen/dl.html) to statistically test for DIF. Because we were doing multiple tests (one for each of the 12 items), we used the Bonferroni correction. The corrected alpha level is .05/12 = .004 and the critical value of the chi-square distribution is 11.04. When using this corrected critical value none of the item parameters exhibit DIF.

DISCUSSION

Our results indicated a one-factor solution fit the data when assessing the dimensionality of the SCI, although some previous research (e.g., Chipuer & Pretty, 1999; Peterson et al., 2008) has found a four-factor solution. However, our study differs from these previous studies in several important ways. First, our response options were dichotomous, which required that we use estimation methods such as full information marginal maximum likelihood or weighted least squares to take into

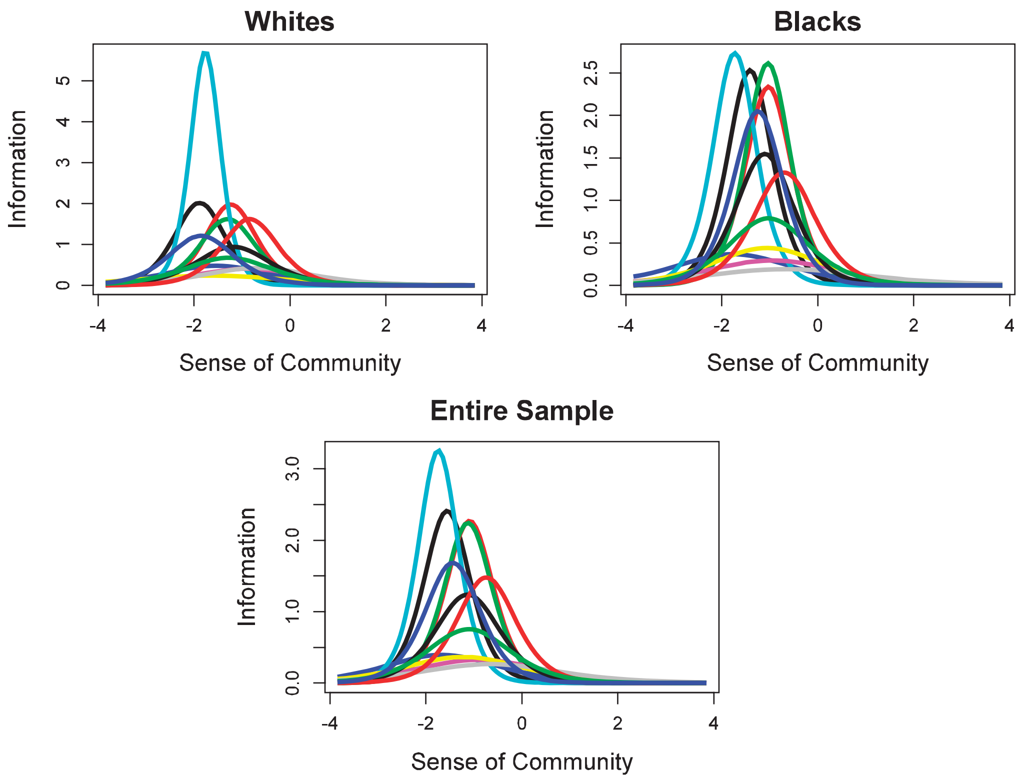

Figure 3. Item information curves by race and for entire sample.

account the non-normally distributed items. Previous research that found four factors has either discounted the non-normally distributed items and fit factor analysis models for normally distributed continuous items (e.g., Chipuer & Pretty, 1999) or used Likert-type response options on a 5-point scale (Peterson et al., 2008). Factor analysis models for continuous normally distributed response options may not be appropriate for Likert responses of five points or fewer. When product moment correlations are computed from categorical data, the coefficient is not free to vary between −1 and 1 because of the dependence on the marginal distribution of each variable. Therefore, the product moment correlation underestimates the true correlation (Muthén, 1983). Olsson (1979a) showed that the difference between the product moment correlation and polychoric correlation estimates for categorical data were large. Olsson (1979b) further showed that the application of factor analysis to ordered categorical data as though it were continuous may result in incorrect conclusions regarding the number of factors and biased estimates of the factor loadings. However, if the items are approximately normally distributed, simulation studies have shown that factor analysis methods for continuous normally distributed responses may be justified if there are at least five categories of response options. Second, several of the previous studies that found four factors for the SCI did not, in fact, perform a factor analysis. Rather, they performed a principal components analysis (e.g., Chipuer & Pretty, 1999). As has been illustrated numerous times in the methodological literature, principal components analysis and factor analysis are not equivalent (see for example, Preacher & MacCallum, 2003). Finally, one of the studies (Peterson et al., 2008) used a different

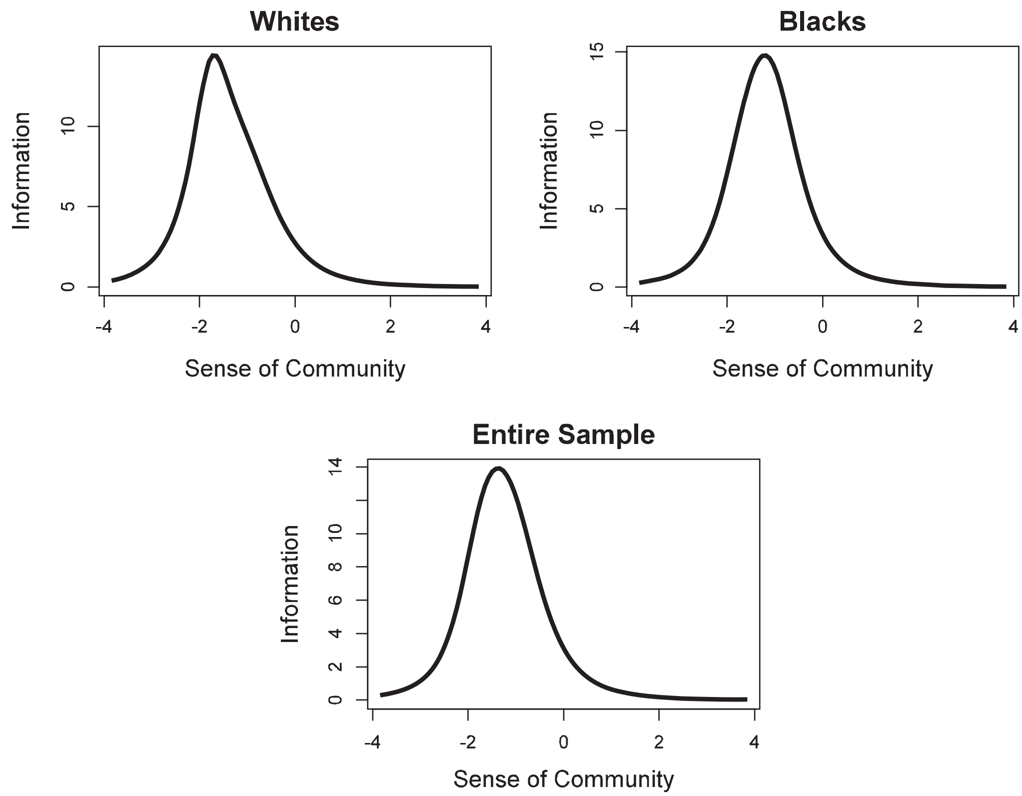

Figure 4. Test information function by race and for entire sample.

scale for SOC than we used here. Therefore, our results are not directly comparable to those of previous research.

Our results illustrate one of the strengths of IRT. In classical reliability analysis as assessed by coefficient alpha, reliability is increased by increasing the number of items in the scale, as long as those items measure the same construct. Further, reliability is required to be the same across the range of the latent construct. In contrast, IRT does not require reliability to be the same across the range of the latent construct. Because of this, even scales with only a few items may be very reliable. However, the items need to provide information across the range of the latent construct. Our results showed that although the reliability as assessed by coefficient alpha was good, the items did not provide information across the entire range of the latent construct. Specifically, the items provided almost no information for individuals who have an above average SOC.

Finally, although we did not find that any of the items exhibited statistically significant DIF, we did find that item 5 was a more reliable measure of SOC for Whites than for Blacks. In other words, item 5 has less measurement error for Whites than for Blacks. When items do not exhibit DIF as in the present study, then the creation of scale scores may proceed as usual. If DIF is detected for a particular item, then the item may be eliminated from the scale, reworded so that it does not exhibit DIF, or item parameters could be separately estimated for the subgroups and these estimates used in creating the scale score. Regardless of whether DIF is present or not, there are numerous methods for creating scale scores based on IRT or simply summing the scores on the item responses. For a review of these methods, see Thissen and Wainer (2001).

Future analyses may be conducted to determine whether there is significant DIF in the SCI for other groups such as those defined by gender, marital status, or education level. In conclusion, we found that that the SCI is invariant across Whites and Blacks and that differences on the SCI may be attributable to true differences in SOC between Blacks and Whites rather than DIF.

REFERENCES

Akaike, H. (1987). Factor analysis and AIC. Psychometrika, 52, 317–332.

BeLue, R., Taylor-Richardson, K.D., Lin, J., McClellan, L., & Hargreaves, M.K. (2006). Racial disparities in sense of community and health status: Implications in community based participatory research interventions targeting chronic disease in African Americans. Journal of Ambulatory Care Management, 29(2), 112–124.

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. Psychometrika, 46, 443–449.

Bock, R.D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. Applied Psychological Measurement, 12, 261–280.

Brodsky, A.E., O'Campo, P.J., & Aronson, R.E. (1999). PSOC in community context: Multi-level correlates of a measure of psychological sense of community in low-income, urban neighborhoods. Journal of Community Psychology, 27(6), 659–667.

Chavis, D., & Pretty, G. (1999). Sense of community: Advances in measurement and application. Journal of Community Psychology, 27(6), 635–642.

Chavis, D.M., & Wandersman A. (1990). Sense of community in the urban environment: A catalyst for participation and community development. American Journal of Community Psychology, 18, 55–81.

Chavis, D.M., Hogge, J.H., McMillan, D.W., & Wandersman, A. (1986). Sense of community through Brunswick's lens: A first look. Journal of Community Psychology, 14(1), 24–40.

Chipuer, H., & Pretty, G. (1999). A review of the sense of community index: Current uses, factor structure, reliability and further development. Journal of Community Psychology, 27(6), 643–659.

Clarke, D.E., Colantonio, A., Rhodes, A.E., & Escobar, M. (2008). Pathways to suicidality across ethnic groups in Canadian adults: The possible role of social stress. Psychological Medicine, 38(3), 419–431.

Cole, S.R. (1999). Assessment of differential item functioning in the perceived stress index-10. Journal of Epidemiology and Community Health, 53, 319–320.

Faircloth, B.S., & Hamm, J.V. (2005). Sense of belonging among high school students representing 4 ethnic groups. Journal of Youth and Adolescence, 34(4), 293–330.

Klein, K.J., & D'Aunno, T.A. (2006). Psychological sense of community in the workplace. Journal of Community Psychology, 14(4), 365–377.

McMillan, D.W. (1996). Sense of community. Journal of Community Psychology, 24(4), 315–325.

McMillan, D.W., & Chavis, D.M. (1986). Sense of community: A definition and theory. Journal of Community Psychology, 14, 6–23.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. Psychometrika, 58(4), 525–543.

Muthén, B. (1983). Latent variable structural equation modeling with categorical data. Journal of Econometrics, 22, 48–65.

Olsson, U. (1979a). Maximum likelihood estimation of the polychoric correlation coefficient. Psychometrika, 44, 443–460.

Olsson, U. (1979b). On the robustness of factor analysis against crude classification of the observations. Multivariate Behavioral Research, 14, 485–500.

Peterson, N.A., Speer, P.W., & McMillan, D.W. (2008). Validation of a brief sense of community scale: Confirmation of the principal theory of sense of community. Journal of Community Psychology, 36(1), 61–73.

Preacher, K.J., & MacCallum, R.C. (2003). Repairing Tom Swift's electric factor analysis machine. Understanding Statistics, 2, 13–32.

Ramírez, M., Ford, M.E., Stewart, A.L., & Teresi, J.A. (2005). Measurement issues in health disparities research. Health Services Research, 40, 1640–1657.

Rizopoulos, D. (2006). LTM: An R package for latent variable modeling and item response theory analyses. Journal of Statistical Software, 17, 1–25.

Sarason, S.B. (1974). The psychological sense of community: Prospects for a community psychology. San Francisco: Jossey-Bass.

Sarason, S.B. (1986). Commentary: The emergence of a conceptual center. Journal of Community Psychology, 14, 405–407.

Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6, 461–464.

Takane, Y., & DeLeeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. Psychometrika, 52, 393–408.

Thissen, D., & Steinberg, L. (1988). Data analysis using item response theory. Psychological Bulletin, 104, 385–395.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group mean differences: The concept of item bias. Psychological Bulletin, 99, 118–128.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), Test validity (pp. 147–169). Hillsdale, NJ: LEA.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 67–113). Hillsdale, NJ: LEA.

Thissen, D., & Wainer, H. (Eds.). (2001). Test scoring. Mahwah, NJ: LEA.